

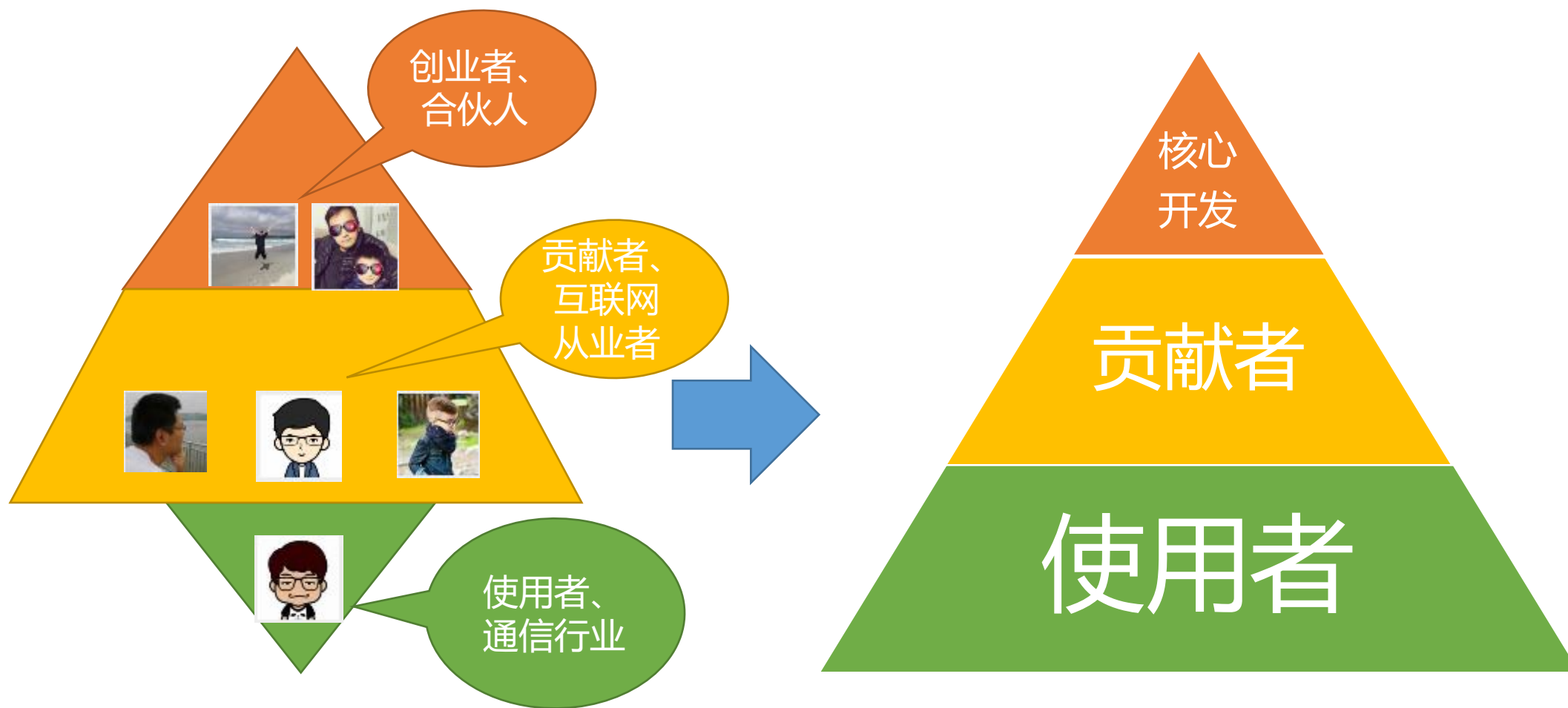
Apache Kylin在电信运营商的 应用案例分享

北京移动 赵磊
zhl@bj.chinamobile.com

赵磊

- 北京移动运维部门大数据团队负责人
- zhl@bj.chinamobile.com
- 13901287305







目录

why



为什么选择麒麟？

doing



Kylin在运营商数据中的应用案例

future



下一步的规划

开源项目的正确打开方式

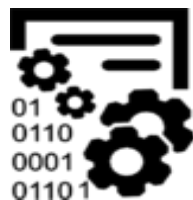
- I. 选
- II. 用
- III. 修改



一、数据规模



用户超过
2000万



原始数据超过
300亿/天



ETL入库
3TB/天



集群规模
20+
400TB

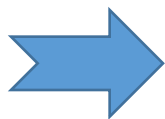
JOB

任务规模超过
800/天

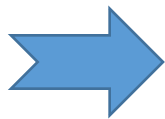
二、数据需求的困境



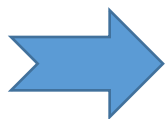
数据的爆炸式增长
探索性数据分析需求旺盛



固定化场景



实时性要求
不高的场景



实时性、灵活
性高的场景

二、解决困境的选择

I. 部署速度快



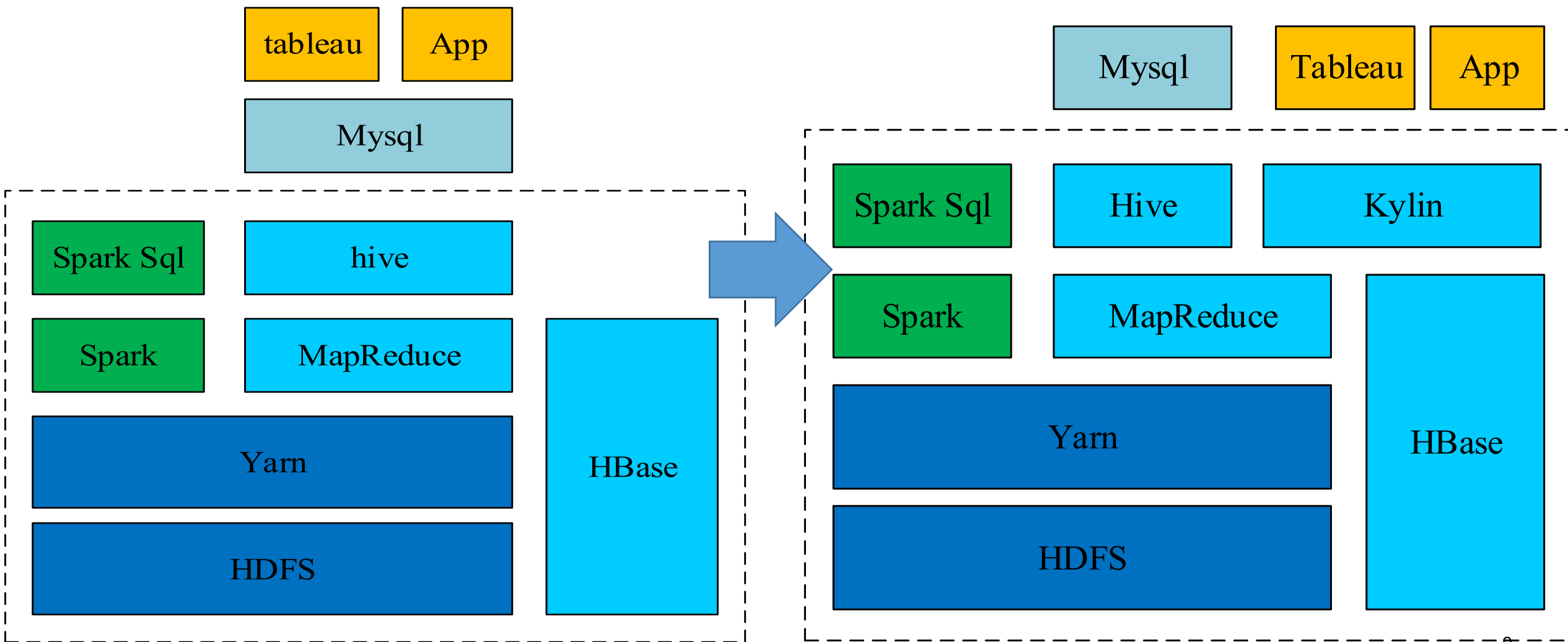
II. 查询速度快

	执行资源	执行时长	备注
hive	86vcores+380GBMEM	1522秒	orc+zlib
spark sql	131vcores+912GBMEM	125秒	orc+zlib
kylin	Hbase5台节点	3.43秒	

*执行测试语句：select rat,count(distinct msisdn) from phone_usertmp where reportdate='20160225' group by rat;

*原始数据大小103GB，条目数11亿

一、离线计算平台的架构的变化



一、应用场景一

ID	终端制式	域名	网络类型	应用类型	应用名称	次数	流量	时长	日期	小时
----	------	----	------	------	------	----	----	----	----	----

I. 统计报表

Dimension：终端制式，域名，网络类型，应用类型，应用名称，日期，小时

Measure：次数求和，流量求和，时长求和，ID排重求和

II. 详单数据

Dimension：ID，终端制式，域名，网络类型，应用类型，应用名称，日期，小时（mandatory=Y）

Measure：次数求和，流量求和，时长求和

原始数据47GB：Cube1：**80分钟（非独占），17GB 膨胀率 36%**

Cube2：**51分钟（非独占），22GB 膨胀率 47%**

二、应用场景二

										WSC
			BJIDCFLA	BJCMCDND	BJCMCDN	BJCMCDNL	BJCMCACHE	BJCMCACHE	BJCMZHILIA	MFLA
APPTYPE	APPNAME	HOSTNAME	G	LFLAG	WSFLAG	XFLAG	HWFLAG	KWFLAG	NFLAG	G
WSCMZHILI	OTHERFL				TIMEDELAY				DNSBJCMC	
ANFLAG	AG	BDRATE	BWRATE	TIMEDELAY	FLAG	SUCRATE	LOADDATE	DNSIDC	DNDL	
DNSBJCMC	DNSBJCM	DNSBJCMC	BJCMCA	DNSBJCMC	DNSBJCMC	DNSBJCMZ		DNSWSCMZ		
DNWS	CDNLX	ACHEWX	CHEWX	ACHEHW	ACHEKW	HILIAN	DNSWSCM	HILIAN	DNSOTHER	
									

*部分字段取值范围非常离散；hostname超过500万，各类*rate的取值范围是0.00-100%

I. 单条查询 < 0.5S

II. 精准查询 < 20S

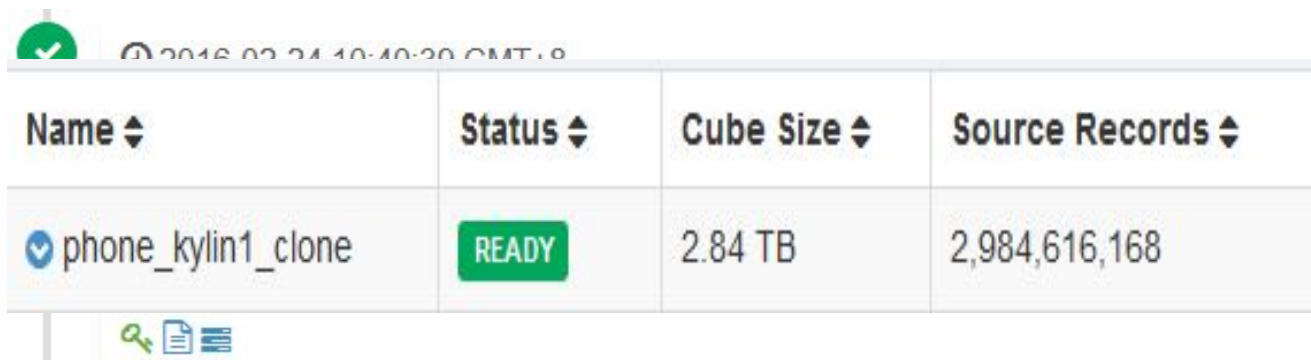
III. 范围查询 > 200S

三、一些注意事项

I. 设计好你的原始数据表

Too many digits for NumberDictionary: 39.8399999999999996. Expect 16 digits before decimal point at max.

II. 设计好你的cube-真的所有的维度都需要吗？



Name	Status	Cube Size	Source Records
phone_kylin1_clone	READY	2.84 TB	2,984,616,168

III. 选择合适维度类型

IV. 理解每一个参数信息

一、升级新版本kylin，实现topN功能

Query String ▾

Start Time: Thu Mar 03 2016 16:31:21 GMT+0800 (UTC+8) -28800000 GMT+8

Rerun Save

Status: **Failed**

Project: phone_test

Cubes:

Results

Scan row count exceeded threshold: 782800, please add filter condition to narrow down backend scan range, like where clause. while executing SQL:

"select host,sum(flow) fl,sum(duration),count(distinct(msisdn)) from phone.phone_usertmp group by host order by fl desc limit 40"

iosapps.itunes....	4491.4844891...	9546834.0	
101.251.217.210	63.655784575...	2844605.0	40274

二、cubing的引擎选择

- I. MapReduce or Spark

三、设计符合需求的拖曳前台界面

- I. 支撑探索性数据查询
- II. 屏蔽后台细节，避免不必要的查询

四、跟进kylin的动态变化

- I. 宽表数据的OLAP查询（维度数远远大于15）
- II. 用户标签快速查询的应用场景（查询条件多变）



中国移动
China Mobile



תודה
Dankie Gracias
Спасибо شكري
Merci Takk
Köszönjük Terima kasih
Grazie Dziękujemy Děkojame
Ďakujeme Vielen Dank Paldies
Kiitos Täname teid 谢谢
Thank You Tak
感謝您 Obrigado Teşekkür Ederiz
감사합니다
Σας ευχαριστούμε
Bedankt Děkujeme vám
ありがとうございます
Tack

中国移动内部资料，
未经允许不得复制、转发、传播。